Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

- ✓ Fall season seems to have attracted more booking. Booking count has significantly increased for each season in 2019 compared to 2018.
- ✓ Clear weather attracted more booking which seems obvious. And in comparison to previous year, i.e 2018, booking increased for each weather situation in 2019.
- ✓ When its not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- ✓ Booking seemed to be almost equal either on working day or non-working day. But, the count increased from 2018 to 2019.
- ✓ 2019 attracted more number of booking from the previous year which was evident from the previous plots.
- ✓ Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year. Number of booking for each month seems to have increased from 2018 to 2019.
- ✓ Wed, Thu, Fri and Sat have more number of bookings as compared to the other days (sun, mon & tue) of the week, for the year 2019.

2. Why is it important to use drop_first=True during dummy variable creation?

Answer:

The intention behind the dummy variable is that for a categorical variable with ‘n’ levels, you create ‘n-1’ new columns each indicating whether that level exists or not using a zero or one. Hence, drop_first=True is used so that the resultant can match up n-1 levels. Hence, it reduces the correlation among the dummy variables.
Eg: If there are 3 levels, the drop_first=True will drop the first column.

When creating dummy variables for categorical data in regression analysis, it is important to use drop_first=True in order to avoid multicollinearity problems known as the "dummy variable trap”. The dummy variable trap occurs when the model has perfect multicollinearity, which means that one of the dummy variables can be perfectly predicted by the others.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:

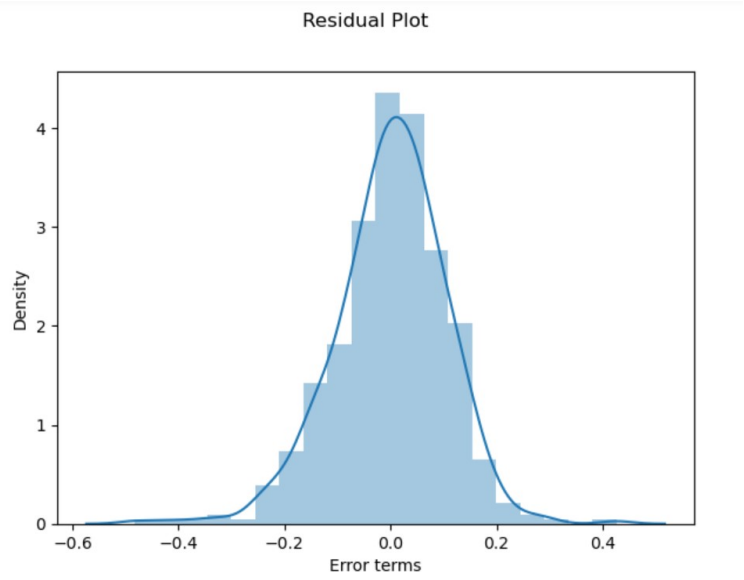
‘temp’ variable has the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

Please find below details for assumptions related to the LR Model:

✓ **Normality of error terms:** Error terms should be normally distributed.



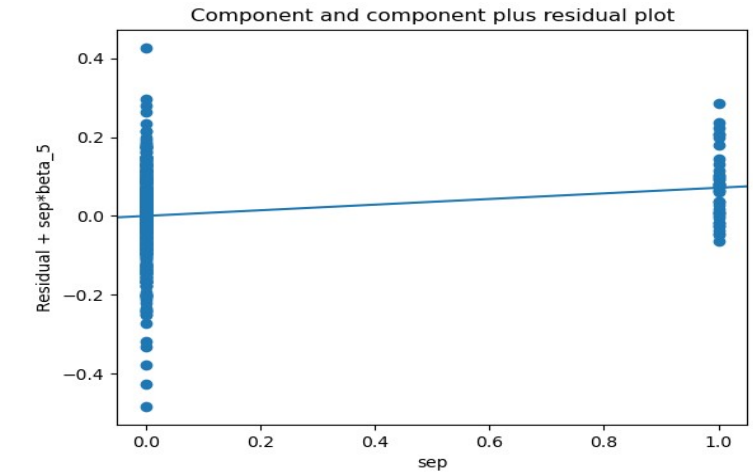
The error terms are fairly normally distributed with mean centered around zero

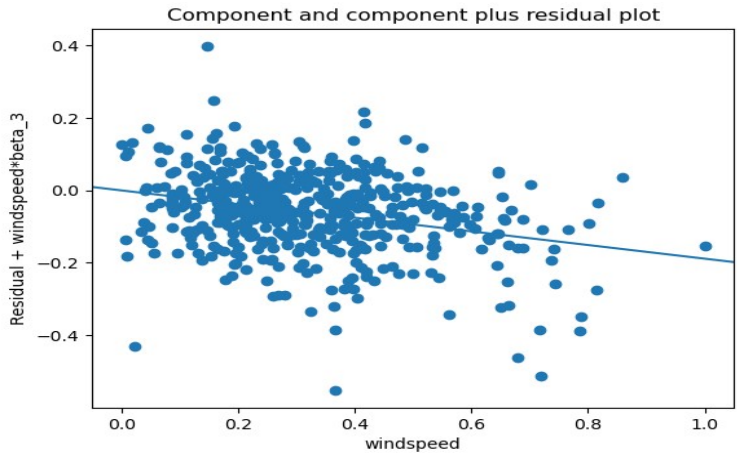
✓ **Multicollinearity check:** There should be insignificant multicollinearity among variables.

	Features	VIF
2	windspeed	4.04
1	workingday	3.29
8	spring	2.65
9	summer	2.00
0	year	1.88
10	winter	1.73
3	jan	1.60
7	Misty	1.57
5	sat	1.56
4	sep	1.18
6	Light_snowrain	1.08

The details of VIF shows no multicollinearity.

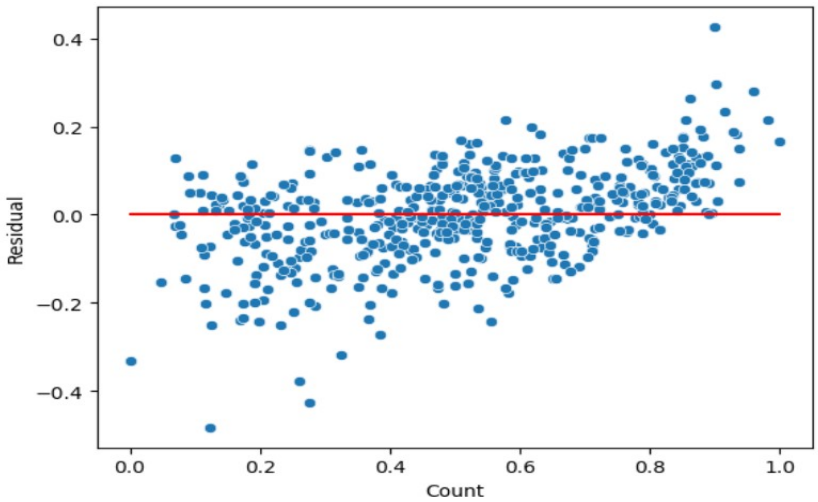
✓ **Linear relationship validation:** Linearity should be visible among variables





Linearity can be observed from the above plots.

✓ **Homoscedasticity:** There should be no visible pattern in residual values.



No visible pattern observed from the above plot for residuals.

✓ **Independence of residuals:**

No auto-correlation.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Answer:

- ✓ sep
- ✓ sat
- ✓ Workingday

General Subjective Questions

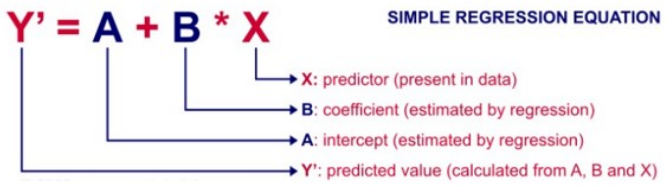
1. **Explain the linear regression algorithm in detail.**

Answer:

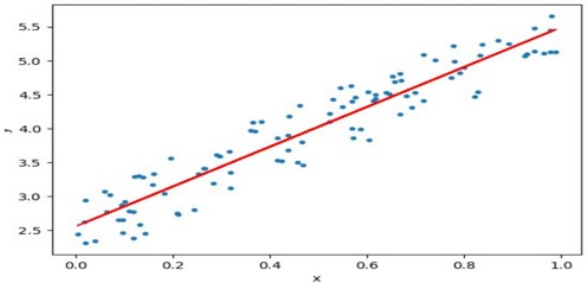
Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single independent variable (x), such linear regression is called simple linear regression. And if there is more than one independent variable, such linear regression is called multiple linear regression.

Assumptions of Linear Regression:

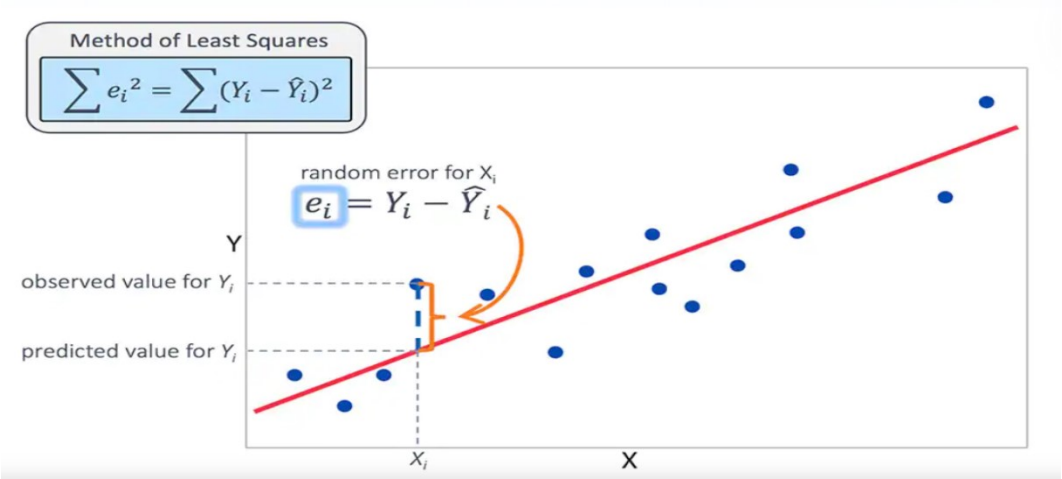
- ✓ **Linearity:** The relationship between the independent and dependent variables is linear.
- ✓ **Independence:** The residuals (errors) are independent.
- ✓ **Homoscedasticity:** The residuals have constant variance.
- ✓ **Normality:** The residuals are normally distributed.
- ✓ **No Multicollinearity:** The independent variables are not highly correlated.



In the figure below, we can see that we have two variables x and y which have some scattered. Through the scatter plot, there is a line passing through. This called a **Regression line** or the **best fit line**.



- The method of least squares is used to find the best-fitting line for the observed data.
- The estimated least squares regression equation has the minimum sum of squared errors, or deviations, between the fitted line and the observations.
- The line of best fit is calculated by using the cost function — Least Sum of Squares of Errors.



The objective of linear regression is to minimize the sum of the squared differences between the observed values and the predicted values (residuals). This method is known as Ordinary Least Squares (OLS).

Steps in Linear Regression:

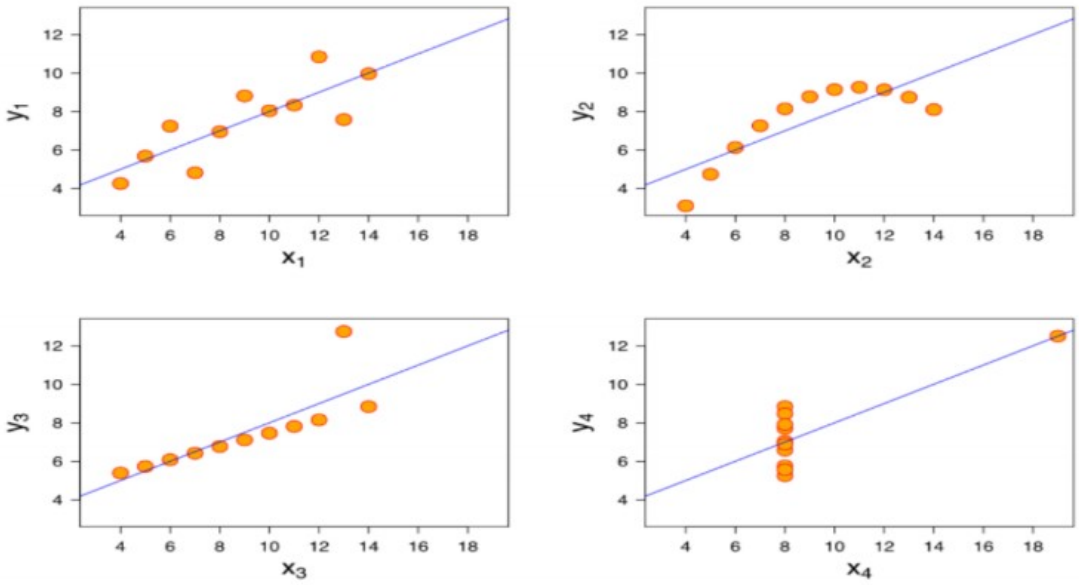
- ✓ **Data Collection and Preparation:** Gather and clean the data. Split the data into training and testing sets.
- ✓ **Exploratory Data Analysis (EDA):** Visualize the data to understand relationships between variables. Check for linearity, correlation, and outliers.
- ✓ **Feature Selection and Engineering:** Select relevant features that influence the dependent variable. Create new features if necessary.
- ✓ **Model Training:** Use the training data to fit the linear regression model. Calculate the coefficients (β0, β1, …, βn).
- ✓ **Model Evaluation:** Evaluate the model using the testing data. Common metrics: R-squared (R2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE).
- ✓ **Prediction:** Use the trained model to make predictions on new data.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the

regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



- 1st data set fits linear regression model as it seems to be linear relationship between X and y.
- 2nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.
- 3rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
- 4th data set has a high leverage point means it produces a high correlation coeff.

Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model.

Importance of Anscombe's Quartet:

Anscombe's quartet demonstrates several important points:

- **Graphical Analysis:** Summary statistics alone can be misleading. Visualizing data is crucial for identifying patterns, relationships, and anomalies that statistics might miss.
- **Outliers Impact:** Outliers can significantly affect regression models and correlations, highlighting the need to carefully inspect and handle them.
- **Model Fit:** The appropriateness of a model cannot be determined by summary statistics alone. It is essential to visualize data to ensure the chosen model fits well.

By understanding and applying the lessons from Anscombe's quartet, analysts can avoid common pitfalls in data analysis and ensure more accurate and insightful conclusions.

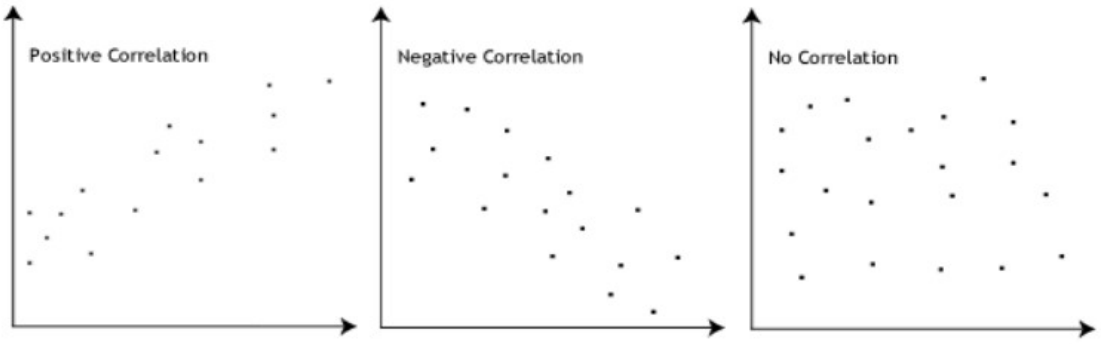
3. What is Pearson's R?

Answer:

Pearson's R is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, R, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

This is shown in the diagram below:



The formula of Pearson's R is:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- x_i and y_i are the individual sample points.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, respectively.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Scaling is the process of adjusting the range of feature values in a dataset so that they fit within a specific range, often to improve the performance and training stability of machine learning algorithms.

Why Scaling is Performed

- **Improves Model Performance:** Many machine learning algorithms perform better when features are on a similar scale. For example, gradient descent converges faster when features are scaled.
- **Enhances Model Interpretability:** Scaling can help make the features more interpretable and comparable, especially when they are on different scales.
- **Prevents Bias Towards Features:** Algorithms that compute distances between data points (e.g., k-NN, SVM or that involve regularization (e.g., linear regression with L2 regularization) can be biased by the scale of features.

Key differences between Normalisation and Standardisation:

Aspect	Normalized Scaling	Standardized Scaling
Formula	$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$	$x' = \frac{x - \mu}{\sigma}$
Range	Typically [0, 1] or [-1, 1]	Mean = 0, Standard Deviation = 1
Sensitivity to Outliers	High	Moderate
Use Case	Non-Gaussian distributions, bounded range	Gaussian distributions, many ML algorithms
Effect	Rescales data to a fixed range	Centers data around mean with unit variance

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

$$\text{VIF}(X_i) = \frac{1}{1-R_i^2}$$

Where R_i^2 is the coefficient of determination of the regression of X_i on all other predictors. If $R_i^2 = 1$ (indicating perfect multicollinearity), then:

$$\text{VIF}(X_i) = \frac{1}{1-1} = \frac{1}{0} = \infty$$

Causes of Perfect Multicollinearity:

- **Duplicate Variables:** Including the same variable more than once in the regression model.
- **Linear Dependence:** One variable is a perfect linear function of another, such as:

  Sum of variables (e.g., X3 = X1+X2).

  Constant multiples (e.g., X2 = 2X1)
- **Dummy Variable Trap:** Including all dummy variables for a categorical feature without dropping one (when using one-hot encoding), causing perfect multicollinearity.

Handling Infinite VIF:

- **Remove One of the Collinear Variables:** Identify and remove one of the perfectly collinear variables.
- **Combine Collinear Variables:** Combine collinear variables into a single feature if they convey the same information.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

The quantile-quantile (Q-Q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.